Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias

William Emigh
Susan C. Herring
Indiana University Bloomington
{emigh,herring}@indiana.edu

Abstract

This paper presents the results of a genre analysis of two web-based collaborative authoring environments, Wikipedia and Everything2, both of which are intended as repositories of encyclopedic knowledge and are open to contributions from the public. Using corpus linguistic methods and factor analysis of word counts for features of formality and informality, we show that the greater the degree of post-production editorial control afforded by the system, the more formal and standardized the language of the collaboratively-authored documents becomes, analogous to that found in traditional print encyclopedias. Paradoxically, users who faithfully appropriate such systems create homogeneous entries, at odds with the goal of open-access authoring environments to create diverse content. The findings shed light on how users, acting through mechanisms provided by the system, can shape (or not) features of content in particular ways. We conclude by identifying sub-genres of web-based collaborative authoring environments based on their technical affordances.

1. Introduction

More than a decade ago, Yates and Orlikowski (1992), drawing on the structuration approach of sociologist Anthony Giddens (1984), observed that human communicators, through their patterns of use grounded in recurring situations, shape the characteristics of communicative genres over time. Yates and Orlikowski simultaneously noted that the medium with which a genre is conventionally associated (for example, email for contemporary organizational memoranda) may imbue communication in that genre with certain structural properties (formatting, stylistic features, etc.). At present, it is widely accepted that these two forces interact: technical specifications predispose users toward certain communicative choices, social dynamics, and normative outcomes, which in turn enable them to realize their situationally-grounded goals (e.g., DeSanctis & Poole, 1994). How this interaction plays out in emergent digital genres, however, remains a question of considerable theoretical and practical interest.

Specifically, the interaction between user choices and system features has implications for various projects currently underway to create repositories of encyclopedic knowledge on the World Wide Web. The encyclopedia, in the sense of "[a] work that contains information on all branches of knowledge or treats comprehensively a particular branch of knowledge[,] usually in articles arranged…by subject,"[1] is a genre that has traditionally taken the form of a print book or books, written by authoritative experts under editorial oversight. In recent years, however, a number of print encyclopedias have been made available in digital form on the web (e.g., the *Encyclopedia Britannica* at www.britannica.com; the *Columbia Encyclopedia* at www.bartleby.com/65/). Other projects have sought to capitalize on the potential of the Internet to bring together diverse expertise rapidly and inexpensively (Sproull & Kiesler, 1991) in order to create general repositories of knowledge that are indigenous to the web.

Two examples of this latter trend are Wikipedia and Everything2. Wikipedia is a wiki authoring environment designed for the purpose of creating a user-written encyclopedia containing information on all subjects. Everything2 is a web-based community bulletin board designed to create, organize and store information about "everything." A question of general interest is the extent to which such user-created online knowledge repositories are similar to, or differ from, expert-created print encyclopedias. In the terminology of Crowston and Williams (2000), do online encyclopedias 'reproduce' their print antecedents, or are they shaped into new forms by the constraints and affordances of the digital medium?

Two prima facie differences between online encyclopedias and traditional print encyclopedias are especially relevant to the present study. First, while content is created by an expert elite for print encyclopedias, online repositories such as Wikipedia and Everything2 are democratic, allowing anyone with access to the Web to contribute. As stated on wiki.org, "Allowing everyday users to create and edit any page in a Web site is exciting in that it encourages democratic use

[1] *Merriam-Webster Online Dictionary*, retrieved September 12, 2004 from http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=encyclopedia&x=18&y=9

0-7695-2268-8/05/$20.00 (C) 2005 IEEE

of the Web and promotes content composition by nontechnical users."[2] At the same time, individuals' writing ability and levels of knowledge vary greatly. Computer-supported collaborative authoring environments thus face a greater challenge than traditional print publications in maintaining a consistent quality of written output (Glover & Hirst, 1995). This challenge is compounded in the case of open-access knowledge repositories, where potentially undesirable authors can contribute as easily as "good" authors.

Overall standards of appropriateness, accuracy and clarity must be maintained if the contents of the site are to have value. Moreover, vandals must be prevented from abusing the rules and resources of the environment (what DeSanctis & Poole, 1994 term 'ironic appropriation') to deface or erase content created by others. Wikis address these concerns by according all users editorial privileges, and by saving cached files of previous content that can be reinstated in case someone erases the entire content of an entry. Community bulletin board sites such as Slashdot and Everything2 employ a reputation system whereby negative ratings on individual entries affect authors' privileges on the site. Underlying both systems is an assumption that good users will collectively enforce standards of quality and consistency. As one wiki commentator notes, "as long as there is a community of well-behaved users prepared to sort things out, problems can be fixed quickly and with little fuss."[3]

These observations give rise to further questions, namely, how similar or different are entries produced in the two types of systems? Which system gives rise to better quality entries? What social processes underlie the production of "good" entries, and how do they shape the conventions of the online encyclopedia genre? Do sites such as Wikipedia and Everything2, which differ in their authoring and editorial mechanisms, produce communicative content that can be characterized as belonging to a single genre?

The present study addresses these questions by comparing the entries produced in Wikipedia with the entries produced in Everything2, focusing on degree of formality in language use. Our findings show that the greater the degree of post-production editorial control afforded by the system, the more formal and standardized the language of the collaboratively-authored documents becomes, analogous to that found in traditional print encyclopedias. Paradoxically, users who faithfully appropriate the Wikipedia system, which affords complete editorial freedom, tend to create homogeneous entries, at odds with the goal of wikis to support the inclusion of diverse voices. The findings shed light on how users, acting through mechanisms provided by the system, can shape (or not) features of content in particular ways. We conclude by identifying sub-genres of web-based collaborative authoring environments based on their technical affordances

2. Background

2.1. Wikis

A wiki is a group communication mechanism invented in 1995 by Ward Cunningham that allows users to create and edit Web page content freely using any Web browser.[4] Two basic criteria make a site a wiki: authorship and version control. In a wiki, all users are potential authors and editors. To modify a node, a user simply clicks on the 'Edit page' link at the bottom of a node, changes the text in a text area, and submits the changes. Input text is converted into HTML by the wiki system.[5] Many wikis allow anyone to modify nodes, although some allow only registered users to do so (it is usually trivial to become a registered user). In order to alleviate the potential problem of "bad" authors, each node has a log of all changes made to it and who made those changes. This makes it easy to revert a node if the content has been deleted or changed.

The system of trust embedded in a wiki is thus primarily social. While the design of a wiki makes it easier to correct data than to add malicious content or delete content (Viégas, Wattenberg, & Dave, 2004), vandals could theoretically prevail through determination and persistence. That they usually do not can be attributed to social factors such as a feeling of community that develops among users, and that gives rise to a sense of responsibility to the site, in part precisely because users have so much power over the content. Some users devote hours each day to monitoring sites, looking out for inaccurate or inappropriate content, and such content is usually removed quickly (Viégas, et al., 2004). Additionally, although the change-logs show who made which changes, the entry itself has no identifying information in it, apart from what the authors insert manually. Anonymous authoring means that the text exists apart from the authors, which may make traditional "flaming" less likely to occur. The fact that wikis succeed as collaborative authoring environments, despite a structure that would appear to encourage widespread abuse, is all the more notable in that the barriers to participation are low.

Most wikis have a specific community purpose (such as FoxPro's wiki, which acts as a forum for FoxPro

[2] http://wiki.org/wiki.cgi?WhatIsWiki

[3] http://www.caslon.com.au/wikiprofile.htm

[4] http://c2.com/cgi/wiki?WikiHistory

[5] Some wikis, including Wikipedia, offer advanced formatting options with a unique syntax. Moreover, although adding or modifying content is quite easy even for non-technical users, refactoring (reorganizing the content of a node and possibly breaking it into sub-nodes) is difficult.

software) and may only be accessible to users on an intranet. The most popular wiki by far (in terms of number of "nodes" or topics) is Wikipedia,[6] which was begun in 2001 by Larry Sanger and Jim Wales, initially to provide a more open alternative to Nupedia, their attempt to create an online encyclopedia with content written by experts (all contributors had to have a Ph.D.).[7] Wikipedia has a separate discussion page associated with each "node" or entry, where contributors can justify and debate the merits of their contributions, but otherwise it resembles other wikis in its technical affordances. While Nupedia's cumbersome editorial model caused production of entries to slow and eventually cease in September 2003, Wikipedia grew rapidly, and as of March 2004, had around 70,000 registered users, of whom 6000 active contributors were working on more than 200,000 articles in English and several hundred thousand in other languages. Its success is also reflected in the fact that it is consulted as a serious information source by many readers, and its entries are cited by mainstream news sources (Lih, 2004).

2.2. Everything2

In 1998, one of the founders of the community weblog Slashdot, Nathan Oostendorp, wrote Everything, a site with the purpose of housing "writings about everything." Everything2[8] is a software upgrade that was originally separate from Everything. The information from the two sites was updated and reincorporated when Everything2 became a single entity in January, 2000.

Like Wikipedia, Everything2 makes it easy for potential authors to contribute. The content for a node is entered in plain-text, which Everything2 converts into HTML. Only registered users are allowed to post content, although anyone may create an account with no verification. Unlike in a wiki, however, only the author of a node can edit that node. This means that content cannot be modified by others directly. Instead, users are explicitly informed of how well they are following social norms by their ranking according to a reputation system.

Everything2 employs an explicit trust metric in which all users have "XP" (eXPerience) that determines their abilities in the system, similar to traditional role-playing games. Beginning authors are unable to rank entries. As they gain XP and write entries, they are given more votes per day. Further experience and entry writing earns them the ability to "cool," or mark as especially interesting, a certain number of entries per day. Authors can gain and lose XP in a variety of ways. Writing a new node gives the author 1 XP. Whenever an established user rates the author's node (either up or down), there is a (random) 1-in-3 chance that the rating will affect the author's XP, and a 1-in-5 chance that the rating will affect the user doing the rating, in the direction of the original rating. This encourages users to give positive feedback more often than negative. A cool gives the author of the cooled node 3 XP and promotes the node to the front page of the Everything2 site. Although all users can see cools, an entry does not show its cumulative rating until it has been rated by the user currently examining it, making cools the only public indication of the popularity of a node. An author may request that a node be removed, e.g., in order to avoid any loss of XP they might incur by having a node that is frequently rated down. In that case, the author loses the 1 XP they got for posting the node.[9]

As in games, the ranking system in Everything2 creates a de facto hierarchy of user privileges, although all users have the same opportunities to earn XP. The editorial infrastructure of Everything2 is also hierarchical: the site administrator appoints editors who have the authority to edit or remove nodes—accompanied by an explanation of why they did so—usually because the nodes are repeatedly negatively evaluated or violate the rules of the site. Bulkeley, Huang, & Lampe (2000, n.p.) note that, "[s]ome users have objected to this system, claiming that it invites abuses, and that views unpopular to this homogenous group will not be able to survive." An example of an edit described by one user as "sucking the personality out of the site" is the removal of profane words. In practice, however, it appears that editors seldom remove user-generated content from the site.

Although it is less widely known than Wikipedia, Everything2 is equally large, with approximately 70,000 registered users, and it also attracts a dedicated community of regular contributors, including some who spend many hours a week on the site and consider it a source of social contacts and emotional support (Bulkeley, et al., 2000).

2.3. Previous research

Very little scholarly research exists on Wikipedia and even less on Everything2. Two recent studies are directly relevant to our questions about the quality and the social processes underlying the creation of content in Wikipedia, however. Lih (2004) compared Wikipedia entries before and after they had been cited in the mainstream press, and found that press citation increased the subsequent "quality" of an entry. In Lih's study, quality was operationalized in terms of the number of edits and the number of unique editors for each node: the more of each, the higher the presumed quality. In March 2004, the average number of edits per topic for all Wikipedia entries was 11.3; of 2,743 active members, 521 "very

[6] http://www.wikipedia.org

[7] http://en.wikipedia.org/wiki/History_of_Wikipedia

[8] http://www.everything2.com

[9] For a discussion of Slashdot's reputation system, see Lampe & Resnick (2004).

active" members contributed 100 edits or more. However, Lih did not analyze the text of the nodes directly, and thus his assumption that more edits and more editors result in higher quality content remains untested.

Viégas, Wattenberg, and Dave (2004) created a visualization tool, *history flow*, to display the dynamic evolution of Wikipedia node content over time. Their application of the tool allows them to identify patterns of vandalism, including mass deletion, offensive copy, phony copy, phony redirection, and idiosyncratic copy. However, most acts of vandalism that occurred during the month of May 2003 were repaired within a matter of minutes by other site members. This rapid "self-healing" is facilitated by a 'recent changes' page on the wiki that lists the latest edits that have been made to the site; Viégas et al. note that some avid members monitor this page closely on a daily basis. As a point of comparison, all content posted to the wiki was found to persist for a median time of 90.4 minutes during the month of May 2003, with less controversial content remaining the longest. Underlying this analysis is a notion of community acceptability of content, rather than quality per se.

Both Lih and Viégas et al. note the importance on Wikipedia of a "neutral point of view" (NPOV), which is promoted explicitly as a mantra of the site. Articles written with a NPOV should "present ideas and facts in such a fashion that both supporters and opponents can agree."[10] Lih likens this policy to that of modern news organizations: "sticking to the facts, attributing sources and maintaining balance" (p. 4). Conciseness is also valued on Wikipedia; Viégas et al. observed that while node size tends to increase over time, 21% of edits reduced the size of a node during the month of edits they analyzed.

Explicit guidelines also exist for how to create a good node on Everything2, in the form of FAQs and node entries. "Noders" are cautioned to avoid "overly subjective" content such as personal lists and political rants, but no particular style is advised, beyond the recommendation to write clearly and "for the ages" (e.g., avoiding current slang). Indeed it is difficult to enforce stylistic norms in Everything2, beyond through the use of the ranking system to "downvote" a poorly-composed entry, although in extreme cases content deemed unacceptable may be removed by the site editors (Bulkeley, et al., 2000). Humor is appreciated in Everything2 nodes, at the same time as noders are advised not to start a node with a humorous definition, at the risk of confusing readers and giving the site a reputation for non-seriousness.

Some readers of Everything2 perceive the quality of the content on the site to be inferior to that produced on wikis, as indicated by the following comment posted on the Everything site:[11]

The biggest problem with Everything is the content. The writers are all trying to be clever, but few of the pages can be taken seriously. So, it is ok for some entertainment, but is not the place to go for enlightenment. Wiki is orders of magnitude better, even though Everything looks flashier. - RalphJohnson

----

(To which an anonymous reader responded: "I disagree. The content on Everything, like on WikiWikiWeb, is exactly what you make it. If you want enlightening content there, type some in.") However, no published study to date has analyzed Everthing2 content, or compared the content produced on Wikipedia and Everything2. The present study aims to fill this gap, with the goal of determining how the different mechanisms for promoting "quality" content on the two sites give rise to characteristic structural and stylistic features.

3. Methodology

3.1. Data

The primary data for this study are the texts of nodes (the equivalent of 'entries' in traditional encyclopedias) common to both Wikipedia and Everything2. Since the contents of both sites are user generated, the nodes are not the same, but there is some overlap. To select the nodes for analysis, we randomly generated a list of 100 nodes from each site, and identified the nodes found on both lists. This resulted in 76 nodes, which we further winnowed to only those containing 100 words or more of text. From these, we selected 15 nodes to represent a range of topic categories, including people (e.g., Karl Marx), places (e.g., Kandahar), things (e.g., pizza), and abstract entities (e.g., corporation), and downloaded the text of those nodes on April 5, 2004.

The extended data include the 30 nodes (15 x 2) from Wikipedia and Everything2, plus analogous content from two additional sources: the 'talk' or discussion pages of Wikipedia, which often accompany a node and provide a forum for contributors to discuss the reasons for their edits to that node, and a traditional print encyclopedia that is available online, the 6th edition of the *Columbia Encyclopedia*. The discussion nodes were added because they are part of the content on the Wikipedia site. The Columbia Encyclopedia entries were added to enable comparison between user-created and traditional (expert-created) encyclopedia content. Nine of the original 15 nodes have discussion pages on Wikipedia, and 10 out of the 15 nodes are included as topic entries in the *Columbia Encyclopedia*, for a total of 49 nodes in the extended data set. Most, albeit not all, of the cognate nodes from the Wikipedia discussion pages and the *Columbia Encyclopedia* contain more than 100 words. The nodes analyzed, and the size of each, are shown for all four

[10] From the Wikipedia guidelines; quoted in Lih (2004).

[11] http://c2.com/cgi/wiki?EverythingAtSlashdot

Table 1. Nodes by source and size (in words)

Node Name	Wikipedia	Everything2	Wikipedia Discussion	Columbia Encyclopedia	Total words
Ben Hogan	594	1547	17	83	2241
British Empire	1518	1301	2493	1625	6937
Corporation	1966	2691	786	762	6205
Fetus	684	172	514	309	1679
Friend[1]	401	936	221	0	1558
Furniture	509	530	0	580	1619
Kandahar	373	2673	0	317	3363
Karl Marx	4547	1927	7906	680	15,060
Mind the Gap[1]	212	1124	449	0	1785
Pizza	1156	2981	291	0	4428
Puffy AmiYumi[1]	234	569	0	0	803
Pulitzer Prize	837	228	0	227	1292
Sing Sing	214	750	0	66	1030
String Theory	573	2371	693	105	3742
VGA[1]	641	286	0	0	927
Total words	14,459	20,086	13,370	4754	52,669

sources in Table 1. Note that Wikipedia forwards from 'Friend' to 'Personal Companion,' but as this is the text that would appear for a user browsing for material on ‘Friend,’ it was included. 'Mind the Gap' refers to warnings delivered to passengers on the British Underground to avoid the gap between the train and the platform, considered by many Americans to be an amusing cultural phenomenon. Puffy AmiYumi is a female Japanese rock band.

3.2. Analytical methods

For the purposes of quantitative analysis, we measured content variability in terms of the degree of formality of the language used in the subject entries (nodes) in the four sources described above. Formality was selected out of all possible properties of the entries because it has been validated in previous studies as an indicator of genre (Biber, 1988, 1995; Heylighen & Dewaele, 1999).

Formality is defined by Heylighen and Dewaele (1999, n.p.) as "expression […] that is context-independent and precise; that is, it represents a clear distinction which is invariant under changes of context." In order to analyze the degree of formality of the entries, we conducted automated frequency counts of words and parts of words identified in previous research as indicating informality or informality in genres of English discourse. To measure the degree of informality of the node text, contractions (*I'm, don't, he's*, etc.) and personal pronouns (*I, we, you, he/she, they* and their case variants) were counted; these have been found to characterize informal genres such as telephone conversations, face-to-face conversations, and personal letters by Biber (1988, 1995) and Heylighen and Dewaele (1999). Formality was measured independently by counting the frequency of common noun-formative suffixes (i.e., *-ment, -(t)ion, -ity, -ism, -ance/ence, -age*), in accordance with the finding of Heylighen and Dewaele that nouns are more frequent in formal genres such as newspapers and scientific writing. The Unix programs 'ptx' and 'grep' were used to count word and suffix frequencies. Node length (in words) and average word length (in letters) were also calculated for each node. Conciseness of message (i.e., communicating more information in fewer words) was found to be a feature of formal, written discourse by Chafe (1982). Short words were found to be characteristic of informal genres in Biber's research.

The resulting counts were subjected to a factor analysis, following the methodology of Biber (1988, 1995), who used factor analysis of frequencies of linguistic features to empirically identify different genres—what he calls 'registers'—of discourse. After computing factor scores from the factor model, we did regression and ANOVA analyses of the factor scores against the source (Wikipedia, Everything2, Wikipedia, Wikipedia Discussion, Columbia) and node (i.e., entry topic) variables in order to identify any significant correlations. The research questions guiding the statistical analyses were: 1) How does the level of formality/ informality of the content of the four sources differ, if at all? and 2) What additional factors, if any, help to explain variations in the data?

Formality is a feature of style, rather than of substance. To arrive at a richer characterization of the content in each source, the quantitative results were supplemented with qualitative observations of the kinds of information provided in the entries, and how the entries were organized. We did not attempt to evaluate the accuracy of, or themes contained in, the content of the entries in this study.

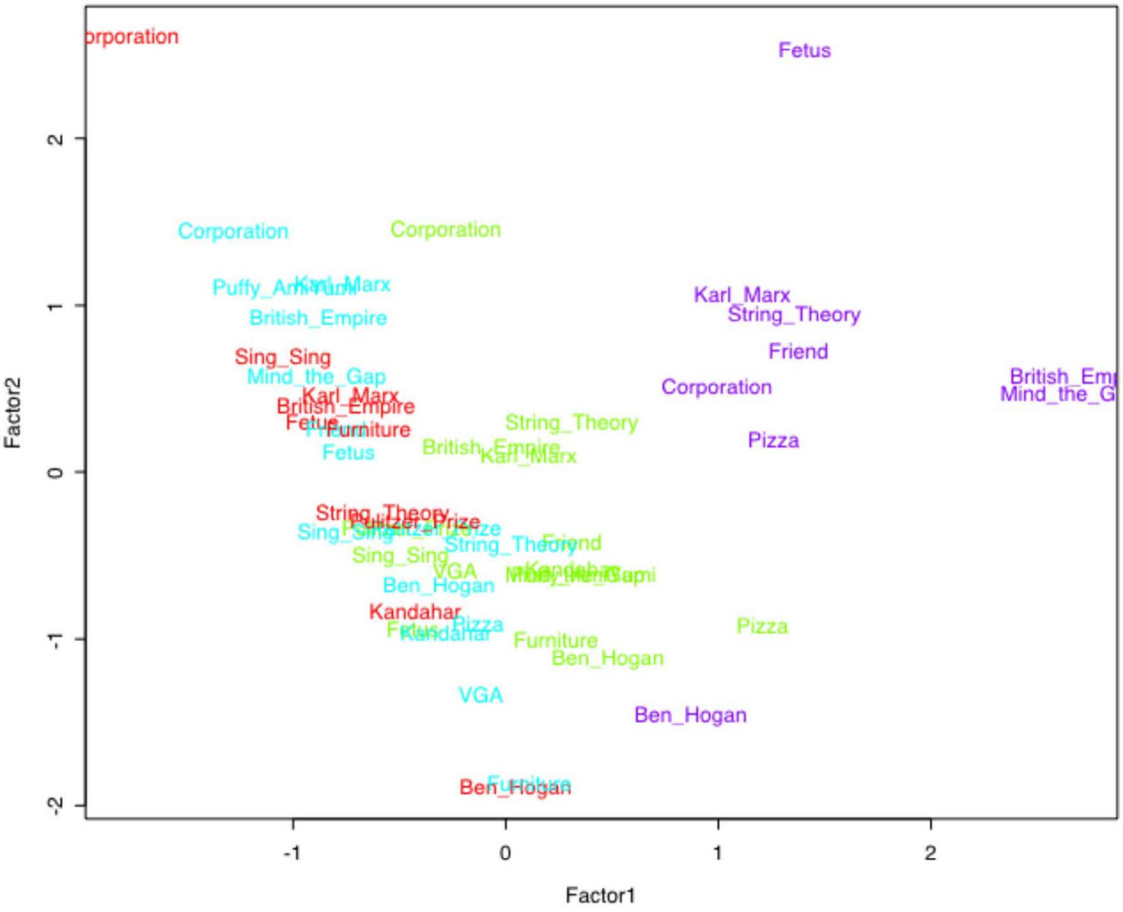

Figure 1. Factor scores of the 49 nodes for factors 1 and 2 (color indicates source; from left to right: red=Columbia, cyan=Wikipedia, green=Everything2, purple=W. Discussion)

4. Findings

4.1. Quantitative results

Simple averages reveal differences in entry length and word length according to source, as shown in Table 2.

Table 2. Entry length and word length

	W	E2	WD	C
Avg. entry (words)	963.93	1339.07	1485.55	475.40
Avg. word (letters)	5.04	4.78	4.86	5.28

The sources can be arranged along a rough continuum on the basis of these results, with the Wikipedia Discussion and Everything2 having the longest entries and the shortest words, the *Columbia Encyclopedia* having the shortest entries and the longest words, and Wikipedia falling in the middle on both counts. To the extent that entry length and word length are indicators of (in)formality, these results suggest that the language of Everything2 (and the Wikipedia Discussion) is more informal than that of Wikipedia, which in turn is less formal than that of the *Columbia Encyclopedia*.

A more precise articulation of this continuum emerges from the factor analysis results, which include the explicit formality and informality features. The factor scores for the 49 nodes in the dataset are displayed as a scatter plot in Figure 1. In this display, a different color indicates nodes from each source.

The analysis identified a two-factor model, in which Factor 1 accounts for 15% of the variation, and Factor 2 accounts for an additional 10% of the variation. Factor 1 has positive loadings for number of words (long entries), contractions, 1st and 2nd person pronouns, and negative loadings for the suffixes *-ment, -(t)ion, -ity*, and *-age*; as such, it is readily interpretable as a dimension of (in)formality. Factor 2 has positive loadings for number of words, 1st person and 3rd person plural pronouns, all of the nominalizing suffixes except *-ity*, and negative loadings for 3rd person singular pronouns. Although this constellation of features is less readily interpretable, and Factor 2 does not achieve statistical significance, the scatter of nodes in Figure 1 suggests that Factor 2 is trying to differentiate among sub-types of nodes, with 'corporation' representing one extreme, and 'Ben Hogan' representing the other extreme, of the dimension.

These observations are further supported by the results of ANOVA analyses of factor scores against the source and node variables. For Factor 1, both source (F=41.0508 on 3 DF, p < .001) and node (F=41.0508 on 14 DF, p < .01) are significant, although upon closer inspection, it emerges that the only significant node is 'corporation', which may not be particularly meaningful in that the entries on 'corporation' in all four sources tended to use a high number of *-tion* suffixes, including in the word 'corporation' itself. More interesting is the finding that Everything2 and the Wikipedia Discussion are significantly different from one another and from the other sources along the formality dimension, but Wikipedia and the *Columbia Encyclopedia* are not significantly different from one another. Statistically speaking, the language of the Wikipedia entries is as formal as that in the traditional print encyclopedia.

For Factor 2, node is significant (F=3.9014 on 14 DF, p < 0.001) but not source; a closer inspection reveals that the nodes 'corporation' and 'Karl Marx' are most significantly different, with the node 'Ben Hogan' selected by the model as the point of comparison. This result is evocative, suggesting that if more nodes were included in the dataset, Factor 2 might identify different types of entry content.

4.2. Qualitative observations

Qualitative observations lend support to the finding that there are differences in content presentation on the source sites, even when the entries are on the same topics. In addition to the use of formal language features and the avoidance of informal and colloquial features, Wikipedia entries are stylistically homogenous, typically describe only a single, core sense of an item,[12] and are often presented in a standard format that includes labeled section headings and a table of contents. These effects can be attributed, in part, to the Neutral Point of View policy of the site (Lih, 2004), which prescribes that all entries should follow a single style, and in part, to the norms of conventional print encyclopedias, which Wikipedia effectively emulates.

In contrast, Everything2 entries make use of informal and colloquial language, including humorous and evaluative expressions, and are internally variable. For example, the individual contributions that make up the ‘Pizza’ nodeshell show variation in the number of personal pronouns and contractions, indicating that individual contributors retain their personal writing styles. The substance of the contributions is similarly variable. The first contribution, for example, describes when a person might want a pizza, while the second contribution describes how to make one.

This tendency is even more apparent in the Everything2 ‘Friend’ nodeshell, where individual entries consist of a blank-verse poem describing what a friend would do, a description of the C++ keyword ‘friend’, and several sentences on the Religious Society of Friends. In contrast, the Wikipedia entry describes ‘friend’ only in the sense of ‘personal companion.' The variability in Everything2 can be attributed to the fact that individual entries remain separate on the page, and no one can edit them to make them more stylistically or substantively consistent. Moreover, because the ability to rate entries is limited by XP and there are potential penalties for rating an entry down, the rating system in Everything2 gives only a coarse control over content, with the result that inconsistencies for which it is otherwise not worth rating a user down typically remain.

At the opposite end of the continuum, the Wikipedia Discussions are consistently informal, making use of emoticons and colloquial expressions such as 'ok.' This consistency does not appear to be caused by contributors editing each other's contributions, but rather reflects an online discussion style typical of webboards and other asynchronous discussion forums (cf. Herring, 2001). Wikipedia 'talk' pages resemble discussion forums, with the exception that authorship of a contribution is not indicated unless the author chooses to include an identifier, as in other wiki contexts, and text can be inserted directly within the text of others. Moreover, discussion content is unlike that in main Wikipedia entries: discussion entries tend to be meta-discussions, including encouragements to write on the topic, discussions on the validity of the content, and discussions on possible refactoring, rather than creating content itself. Users appear to employ stylistic means to distinguish discussion text from entries proper.

Finally, we observed variation according to node topic within each source. For example, although both 'Ben Hogan' and 'Karl Marx' are famous individuals (and descriptions of both thus make use of 3rd person singular pronouns), Hogan is typically described through narrative (the golfer is best remembered for making an inspiring comeback after nearly being killed in an automotive accident in 1948), while the description of Marx is interlarded with expository statements of a philosophical, abstract nature (which are more likely to involve nominalized forms such as *-ism, -ment*, etc.). The entry for the abstract entity 'corporation' contains even more expository features. This observation may help to explain

[12] If multiple senses are available, they tend to be split off ('refactored') into separate entries; see also Viégas, et al. (2004).

why these three nodes were identified in the factor analysis as significantly different: Factor 2 may represent a dimension of narrative vs. expository text.

In what follows, we present extracts from the entries for 'friend' and 'string theory' to illustrate these qualitative generalizations. In its entry on 'personal relationships,' to which readers are directed when they search for 'friend,' the Wikipedia definition begins formally, with a sentence devoid of personal pronouns and containing a nominalization ('connection'):

> The phrase **personal relationship** characterises some sort of connection between two or more people. [W]

In contrast, the Wikipedia discussion of this node includes many informal features common to web chats, such as first person pronouns, contractions, emoticons (X_X), and informal lexicon ('info'):

> This page- which "Friendship" redirects to- contains some relevant info, but seems to discuss romantic relationships more than it does normal friendship; there's nothing here on the formation of friendship, what defines a friendship, the typical emotional dependance of humans on friendship, how friendships drift apart, and so forth. I'm sure Wikipedia can do better than this in an issue so fundamental to society. (And I'd try to do something myself, but 1. I'm not sure on whether to edit "Friendship" into its own article or edit this one, and 2. I'm... tired... X_X so this may have to wait a bit.) --AceMyth 01:50, 19 Dec 2003 [WD]

The Everything2 entries for 'friend' fall on both ends of the spectrum, ranging from informal:

> The person who will come all the way across town to the emergency room in which you have been stranded for seven hours.
>
> You need not have called him.
>
> If you did, you will have forgotten to provide the name of the hospital itself, nevermind your own name. [E2]

to a formal entry that is taken directly from an out-of-copyright dictionary:

> One who looks propitiously on a cause, an institution, a project, and the like; a favorer; a promoter; as, a friend to commerce, to poetry, to an institution. [E2]

The *Columbia Encyclopedia* does not include an entry on 'friend,' perhaps because the concept is considered too basic for a traditional encyclopedia. However, for the entry on 'string theory,' similarities in formality between the *Columbia Encyclopedia* and Wikipedia can be seen in the first paragraph from each source:

> [D]escription of elementary particles based on one-dimensional curves, or “strings,” instead of point particles. Superstring theory, which is string theory that contains a kind of symmetry known as supersymmetry, shows promise as a way of unifying the four known fundamental forces of nature. The strings are embedded in a space-time having as many as 10 dimensions—the three ordinary dimensions plus time and seven compactified dimensions. The energy-scale at which the stringlike properties would become evident is so high that it is currently unclear how any of the forms of the theory could be tested. [C]

> A **string theory** is a physical model whose fundamental building blocks are one-dimensional extended objects (strings) rather than the zero-dimensional points (particles) that were the basis of most earlier physics. For this reason, string theories are able to avoid problems associated with the presence of pointlike particles in a physical theory. Detailed study of string theories has revealed that they contain not just strings but other objects, variously including points, membranes, and higher-dimensional objects. As discussed below, it is important to realize that no string theory has yet made firm predictions that would allow it to be experimentally tested. [W]

Although these definitions are worded differently, their substance and style (e.g., nominalizations in -(t)ion, lack of pronouns, abstract rather than human grammatical subjects) are similar. In contrast, the first contribution to the Everything2 nodeshell on 'string theory' is written in a first person, more informal style:

> The best popular book on the topic of String Theory has got to be Brian Greene's "The Elegant Universe." After reading that book, I found that I finally understood quite a bit about what this theory really means.
>
> String Theory, now called Superstring Theory due to its inclusion of supersymmetry, is gradually unifying its varieties of theories on strings into one large theory called "M-Theory". M-Theory uses an 11 dimensional universe, with three extended spatial dimensions and one time dimension. The rest of the dimensions are curled up in a Calibi-Yau shape, which I can't even begin to explain. [E2]

There is no Wikipedia discussion page for the entry on 'string theory.'

These observations indicate that it is not just language style that differs across the four sources, but presentation, consistency, and scope of the content as well. *Columbia Encyclopedia* and Wikipedia entries are systematic, standardized, and narrow in scope; Everything2 entries are variable, polyvocal, and broad in scope; and Wikipedia discussion entries, which contain mostly metacommentary, resemble interactive forms of computer-mediated communication. These findings have implications for the genre classification of Wikipedia and Everything2, as well as for the strengths and weaknesses of different system designs for online knowledge repositories.

5. Discussion

In this study, we have compared the presentation and style of content in entries in two user-created online knowledge repositories with different technical affordances, extending the same methods of analysis to two cognate sources, discussions associated with main entries, and a traditional print encyclopedia available online. The results of the four-way comparison reveal a continuum of formality and standardization, with the traditional encyclopedia and the interactive discussion at opposite extremes. Wikipedia and Everything2 differ significantly from one another, with Wikipedia towards the formal, standardized end, and Everything2 towards the informal, variable end of the continuum. Surprisingly, Wikipedia is statistically indistinguishable from the print encyclopedia in terms of the formality features measured in this study.

These findings suggest that what we have heretofore been considering as the genre of online encyclopedia is not a uniform set of communicative practices. Wikipedia and Everything2 have functional and structural characteristics in common: they aim to be repositories of general knowledge, they are available online, their contents are searchable, their entries make use of hyperlinks, they are created by multiple non-expert authors who form a community around the practice of creating content for the site, and they are consulted (to varying degrees) by Internet users seeking information on a wide range of topics. These commonalities justify considering the two sites as exemplars of a single genre, according to the standard definition of a genre as recurrent communication characterized by a common purpose, structures, and participant roles (cf. Yates & Orlikowski, 1992). At the same time, the mechanisms for editorial control differ; there are differences in the normative guidelines provided on each site (e.g., Wikipedia's Neutral Point of View policy, which is not shared by Everything2); and the entries themselves are stylistically and substantively different. According to Biber (1988, 1995), the statistical identification of a cluster of linguistic features that distinguish one communication type from another constitutes grounds for positing separate genres.

Our solution to this apparent classification paradox is to propose that Wikipedia and Everything2 are both members of the 'online knowledge repository' genre, but that they represent different genres (or sub-types) of online collaborative authoring environments. Wikipedia is part of the world of wikis, which are used not only to create encyclopedias but also collaboratively-authored FAQs and documentation for software products. Everything2 is kin to other collaborative content systems that incorporate reputation metrics, such as Slashdot, Kuro5hin, and Fark. These sub-types follow from the technical affordances of the sites—notably, the mechanisms relating to editorial control. As noted by Yates and Orlikowski (1992), properties of the medium can shape genre conventions; in this case, editorial mechanisms shape characteristics of formality and variability.

It still remains to explain why Wikipedia—a user-created encyclopedia—is largely indistinguishable stylistically from the expert-created *Columbia Encyclopedia*, since the two are produced by radically different technical means. How is it that the wide-open participation structure of a wiki can reproduce traditional print norms? We believe that two social forces play a role in this outcome. First, Wikipedia users appropriate norms and expectations about what an 'encyclopedia' should be, including norms of formality, neutrality, and consistency, from the larger culture (cf. DeSanctis & Poole, 1994). Second, those norms are enforced through the agency of dedicated, socially-approved members of the Wikipedia community. The common structural elements in the Wikipedia entries suggest that one user (or a small group of users) has changed existing nodes to conform to stylistic norms. The “good” users, who as have been noted by Lih (2004) and Viégas, et al. (2004) are extremely active in the system, may also remove experimental content before most users are able to see it. Their level of activity and interest give them more effective control over the system than casual users. This is a case of genre reproduction (Crowston & Williams, 2000). In contrast, Everything2 is a hybrid product of the Web—a blend of discussion forum and knowledge repository—thus arguably 'emergent' in Crowston & Williams' terms.

Ironically, "good" rank-and-file users on Wikipedia achieve in near-absolute terms what some participants in Everything2 fear from self-interested administrators (Bulkeley, et al., 2000), but which Everything2 comes nowhere close to realizing: imposition of stylistic homogeneity. While this could be viewed as an accomplishment—Wikipedia is increasingly being consulted as a standard reference, in part due to its resemblance to traditional print encyclopedias—it is at odds with the goal of the wiki (and user-created content) movement to create content incorporating diverse perspectives, and more generally to foster new and better communication practices. Notably, it suggests that a few active users, when acting in concert with established norms within an open editing system, can achieve ultimate control over the content produced within the system, literally erasing diversity, controversy, and inconsistency, and homo-genizing contributors' voices. This is an unintended, and to our knowledge previously unnoted, side effect of the "democratic" affordances of wikis.

In contrast, Everything2 realizes the goal of diverse content. Even if Everything2 were to adopt a "neutral point of view" policy, stylistic homogeneity is not enforceable on the site. In the Everything2 system, experimental content—providing it does not run afoul of the site editors—may remain up as long as its author wishes, allowing time for a majority of casual users to make up for the ratings of a few very active users. Some regular contributors to Everything2 have more influence than others—the reputation system ensures it—but authors preserve ultimate control over their entries, rendering the site's contents diverse and, at times, "noisy" and subjective.

Which system produces better content? Although this study did not directly investigate quality of content, the results of the analysis suggest that there is no simple answer to this question, but rather that the answer depends on the goals and preferred styles of users. A system that empowers authors to retain their content in the face of diverging views or criticism can result in more varied, original, and personal—albeit less polished or coherent—content. Another web-based authoring system that embraces these values is the weblog, to which Everything2 and its parent Slashdot are related; weblogs are claimed to be especially well-suited to political commentary and grassroots journalism (Lasica, 2001; cf. Herring, Scheidt, Bonus, & Wright, 2004). Conversely, a system that empowers anyone to edit others' content may attract self-appointed norm enforcers, resulting in more literate, concise, and stylistically-consistent—albeit less original and varied—content. Wikis are well-suited to corporate purposes such as creating product manuals, whereas a system like Everything2 might be a better choice for soliciting honest feedback on products. Moreover, Wikipedia's success demonstrates that it meets users' needs for reliable, up-to-date information. Indeed, with its searchable content, convenient online access, and ability to create entries on recent events quickly, Wikipedia improves on traditional information sources, especially for the content areas in which it is strong, such as technology and current events (Lih, 2004). Each system thus has its limits and appropriate uses; an understanding of these can improve the future design and implementation of such systems.

6. Conclusion

In this study we have observed that the technical affordances of online collaborative authoring systems interact with social norms to (re)produce genre structures, consistent with the claims of Giddens' structuration theory as applied to digital environments by Yates and Orlikowski (1992) and DeSanctis and Poole (1994). Moreover, we have proposed that such interactions give rise to genre sub-types, in this case revolving around the distinction between editorial vs. authorial control and its consequences for the style and presentation of encyclopedic content.

Future research might test this proposal by analyzing the evolution of entries in online knowledge repositories over time. If our theory of the impact of "good" users is correct, we might expect to find evidence of increasing formality and homogeneity across the lifespan of a Wikipedia entry, as well as differences in formality between beginning and experienced contributors, but relatively little change across the lifespan of an Everything2 entry. It would also be informative to compare Wikipedia with open-access wikis that lack explicit guidelines for appropriate content, to evaluate the impact of the neutral point of view policy. Investigations of this sort would help to clarify further the effects of social as opposed to technological structures on the conventions of digital genres.

.

7. References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Bulkeley, N., Huang, W., & Lampe, C. (2000). An examination of the Everything system as a tool for building network communities. Unpublished ms, University of Michigan School of Information.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and Written Language: Exploring Orality and Literacy* (pp. 35-53). Norwood, NJ: Ablex.

Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society,* 16 (3), 201-216.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science*, 5 (2), 121-147.

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structure*. Berkeley, CA: University of California Press.

Glover, A., & Hirst, G. (1995). Detecting stylistic inconsistencies in collaborative writing. In T. van der Geest et al. (Eds.), *Writers at work: Professional writing in the computerized environment* (pp. 147-168). London: Springer.

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds), *The Handbook of Discourse Analysis* (pp. 612-634). Oxford: Blackwell Publishers.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. *Proceedings of HICSS-37*. Los Alamitos: IEEE Press.

Heylighen, F., & Dewaele, J-M. (1999). Formality of language: Definition, measurement and behavioral determinants. Internal Report, Center "Leo Apostel", Free University of Brussels.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *CHI 2004*, 543-550.

Lasica, J. D. (2001). Blogging as a form of journalism. *USC Annenberg Online Journalism Review*, May 24. http://www.ojr.org/ojr/workplace/1017958873.php

Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. Paper presented at the 5th International Symposium on Online Journalism, April 16-17, UT Austin.

Sproull, L., & Kiesler, S. (1991). *Connections: New Ways of Working in the Networked Organization*. Cambridge, MA: MIT Press.

Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with *history flow* visualizations. *CHI 2004*, 575-582.

Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review, 17* (2), 299-326.